



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

### An ontology for major histocompatibility restriction

**Citation for published version:**

Vita, R, Overton, JA, Seymour, E, Sidney, J, Kaufman, J, Tallmadge, RL, Ellis, S, Hammond, J, Butcher, GW, Sette, A & Peters, B 2016, 'An ontology for major histocompatibility restriction', *Journal of Biomedical Semantics*, vol. 7, no. 1. <https://doi.org/10.1186/s13326-016-0045-5>

**Digital Object Identifier (DOI):**

[10.1186/s13326-016-0045-5](https://doi.org/10.1186/s13326-016-0045-5)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Publisher's PDF, also known as Version of record

**Published In:**

Journal of Biomedical Semantics

**General rights**

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



DATABASE

Open Access



# An ontology for major histocompatibility restriction

Randi Vita<sup>1\*</sup> , James A. Overton<sup>1</sup>, Emily Seymour<sup>1</sup>, John Sidney<sup>1</sup>, Jim Kaufman<sup>2</sup>, Rebecca L. Tallmadge<sup>3</sup>, Shirley Ellis<sup>4</sup>, John Hammond<sup>4</sup>, Geoff W. Butcher<sup>5</sup>, Alessandro Sette<sup>1</sup> and Bjoern Peters<sup>1</sup>

## Abstract

**Background:** MHC molecules are a highly diverse family of proteins that play a key role in cellular immune recognition. Over time, different techniques and terminologies have been developed to identify the specific type(s) of MHC molecule involved in a specific immune recognition context. No consistent nomenclature exists across different vertebrate species.

**Purpose:** To correctly represent MHC related data in The Immune Epitope Database (IEDB), we built upon a previously established MHC ontology and created an ontology to represent MHC molecules as they relate to immunological experiments.

**Description:** This ontology models MHC protein chains from 16 species, deals with different approaches used to identify MHC, such as direct sequencing versus serotyping, relates engineered MHC molecules to naturally occurring ones, connects genetic loci, alleles, protein chains and multi-chain proteins, and establishes evidence codes for MHC restriction. Where available, this work is based on existing ontologies from the OBO foundry.

**Conclusions:** Overall, representing MHC molecules provides a challenging and practically important test case for ontology building, and could serve as an example of how to integrate other ontology building efforts into web resources.

**Keywords:** Major histocompatibility complex, Ontology, MHC, Immune epitope

## Background

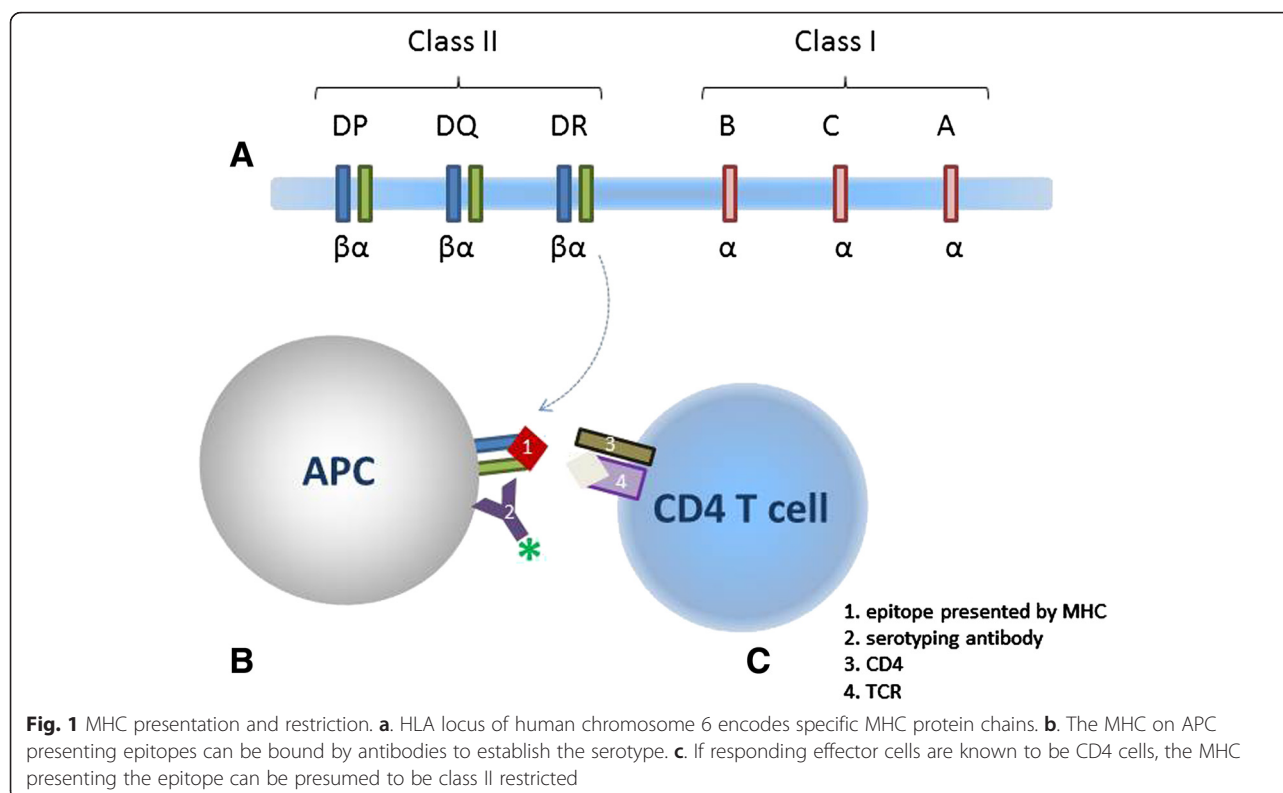
Major histocompatibility complex (MHC) proteins play a central role in the adaptive immune system. First discovered due to their role in transplant rejection, MHC molecules are encoded by a large family of genes with wide variation within each species. MHC molecules typically bind peptide fragments of proteins and display them on the cell surface where they are scanned by T cells of the immune system. If a peptide fragment is displayed by MHC, it can trigger a T cell immune response. Peptides triggering a response are referred to as 'epitopes'. Thus, binding of epitopes to MHC molecules is an integral step for immune recognition. The specific MHC molecule that presents an epitope to a T cell is known as its "MHC restriction", often called its MHC restriction (or restricting) element. Accurately representing

this MHC restriction, which can be determined in different manners, is the goal of the work presented here. Most MHC molecules consist of two protein chains, of which at least one gene is present within the MHC locus. In humans this locus is known as the human leukocyte antigen (HLA) and is depicted in Fig. 1a. There are thousands of different allelic variants of these genes coding for different proteins that result in diverse MHC binding specificities found in the human population. The most precise way of specifying MHC restriction is to identify the exact protein chains that make up the MHC molecule. However, until recently such exact molecular typing was not possible, and patterns of antibody binding were utilized to group MHC molecules together into serotypes that share a common serological (antibody based) recognition pattern, as shown in Fig. 1b. Tying such traditional serotype information together with current sequence based MHC typing techniques is one of the goals of our study. In yet other cases, such as inbred mouse strains, MHC restriction is narrowed down based on the haplotype of the animal, the set of alleles present on

\* Correspondence: rvita@liai.org

<sup>1</sup>La Jolla Institute for Allergy and Immunology, 9420 Athena Circle La Jolla, San Diego, California 92037, USA

Full list of author information is available at the end of the article



a single chromosome and thus expressed consistently together in select subspecies or strains. Another way MHC restriction is sometimes inferred is based on the T cells recognizing the epitope. MHC molecules are divided into three classes: MHC class I, MHC class II, and non-classical MHC. MHC class I molecules present epitopes to CD8<sup>+</sup> T cells and are made up of one alpha chain and one  $\beta$ 2 microglobulin chain, which is invariant and encoded outside the MHC locus. MHC class II molecules present epitopes to CD4<sup>+</sup> T cells and are composed of one alpha and one beta chain, as shown in Fig. 1c. Thus knowing if the responding T cell expresses CD4 versus CD8 can be used to narrow down the possible MHC restriction into classes. At the same time, current research has identified that some T cell populations do not follow this pattern exactly (e.g. some T cells recognizing MHC-II restricted epitopes express CD8). It is therefore important to capture not only the inferred restriction information, but also the evidence upon which it was based.

## Methods

The Immune Epitope Database ([www.iedb.org](http://www.iedb.org)) presents thousands of published experiments describing the recognition of immune epitopes by antibodies, T cells, or MHC molecules [1]. The data contained in the IEDB is primarily derived through manual curation of published literature, but also includes some directly submitted

data, primarily from NIAID funded epitope discovery contracts [2]. The goal of the current work was to represent MHC data as they are utilized by immunologists to meet the needs of the IEDB users. We collected user input at workshops, conferences and the IEDB help system regarding how they wanted to retrieve data from the IEDB regarding MHC restriction. These requests were used to identify goals for this ontology project and the final ontology was evaluated if it could answer these requests. As shown in Additional file 1: Table S1, an example of such a request was to be able to query for epitopes restricted by MHC molecules with serotype 'A2' and retrieve not only serotyped results but also those where the restriction is finer mapped e.g. to MHC molecule A\*02:01 which has serotype A2. We set out to logically represent the relationships between the genes encoding MHC, the haplotypes linking together groups of genes in specific species, and the individual proteins comprising MHC complexes, in order to present immunological data in an exact way and to improve the functionality of our website. Our work builds on MaHCO [3], an ontology for MHC developed for the StemNet project, using the well-established MHC nomenclature resources of the international ImMunoGeneTics information system (IMGT, <http://www.imgt.org>) for human data and The Immuno Polymorphism Database (IPD, <http://www.ebi.ac.uk/ipd>) for non-human species. It contains

118 terms for MHC across human, mouse, and dog. We were encouraged by the success of MaHCO in expressing official nomenclature using logical definitions. However, we needed to extend it for the purpose of the IEDB to include data from a growing list of 16 species, as well as data about MHC protein complexes (not just MHC alleles), haplotypes and serotypes. Thus, our current work goes beyond MaHCO, and we have utilized this opportunity to also enhance the integration with other ontological frameworks.

We used the template feature of the open source ROBOT ontology tool [4] to specify the content of our ontology in a number of tables. Most of the tables correspond to a single “branch” of the ontology hierarchy, in which the classes have a consistent logical structure, e.g. gene loci, protein chains, mutant MHC molecules, haplotypes, etc. The OWL representation of our ontology is generated directly from the tables using ROBOT. This method enforces the ontology design patterns we have chosen for each branch, and makes certain editing tasks easier than with tools such as Protégé.

## Results and discussion

Our MHC Restriction Ontology (MRO) is available in a preliminary state at <https://github.com/IEDB/MRO>. It is based on existing ontology terms, including: ‘material entity’ from the Basic Formal Ontology (BFO) [5], ‘protein complex’ from The Gene Ontology (GO) [6], ‘protein’ from The Protein Ontology (PRO) [7], ‘organism’ from The Ontology for Biomedical Investigations (OBI) [8], ‘genetic locus’ from The Reagent Ontology (REO) [9], ‘has part’, ‘in taxon’, and ‘gene product of’ from The Relation Ontology (RO) [10]. The NCBI Taxonomy was used to refer to each species [11]. Although it is not yet complete, we strive to conform to Open Biological and Biomedical Ontologies (OBO) [12] standards. MRO currently contains 1750 classes and nearly 9000 axioms, including more than 2100 logical axioms. Its DL expressivity is “ALEI”, and the HermiT reasoner [13] completes reasoning in less than 10 seconds on a recent laptop.

Synonyms were also included, as immunologists often utilize synonyms that are either abbreviations or based on previous states of the nomenclature. The current MHC nomenclatures for various species have been revised through several iterations. In order to ensure accuracy and remain up to date with the latest nomenclature, we referred to the well-established MHC nomenclature resources of the IMGT and IPD. For specific species where the literature was most formidable, such as chicken, cattle, and horse, we collaborated with experts in these fields. These experts reviewed the encoded hierarchy by determining whether the inferred parentage hierarchy in their area of expertise reflected their input.

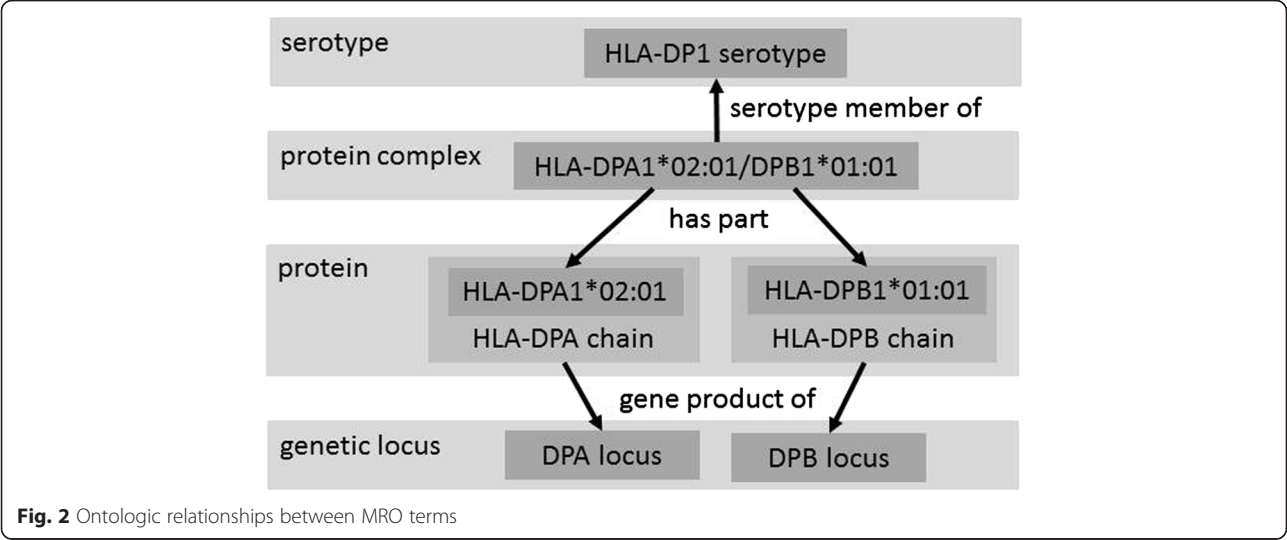
Each MHC molecule for which the IEDB has data is modeled as a protein complex consisting of two chains. Each chain is a gene product of a specific MHC genetic locus. For certain species, sub-loci are also defined, when useful. For example, as shown in Fig. 2 HLA-DPA1\*02:01/DPB1\*01:01 consists of one HLA-DPA1\*02:01 chain, encoded by the DPA sub-locus of DP, and one HLA-DPB1\*01:01 chain, encoded by the DPB1 sub-locus of DP. Together these two chains make up one DPA1\*02:01/DPB1\*01:01 MHC molecule.

When the identity of only a single chain of the complex is known, a “generic” second chain is used to make up the MHC complex. Thus, MHC restriction of HLA-DPB1\*04:02 is modeled as one HLA-DPB1\*04:02 chain in complex with an HLA-DPA chain that is not further specified, as shown within the context of the hierarchy in Fig. 3.

The data in the ontology drives the Allele Finder on the IEDB website, available at <http://goo.gl/r8Tgrz>, an interactive application that allows users to browse MHC restriction data in a hierarchical format. We evaluated the ability of MRO to meet the needs of IEDB users, as shown in Additional file 1: Table S1, and found it to meet our initial goals. Currently the use of the ontology is behind the scenes, but we have requested namespace and permanent identifiers from The Open Biomedical Ontologies (OBO). As soon as these identifiers are in place, they will be utilized and displayed on the IEDB website to allow users to link out to the ontology.

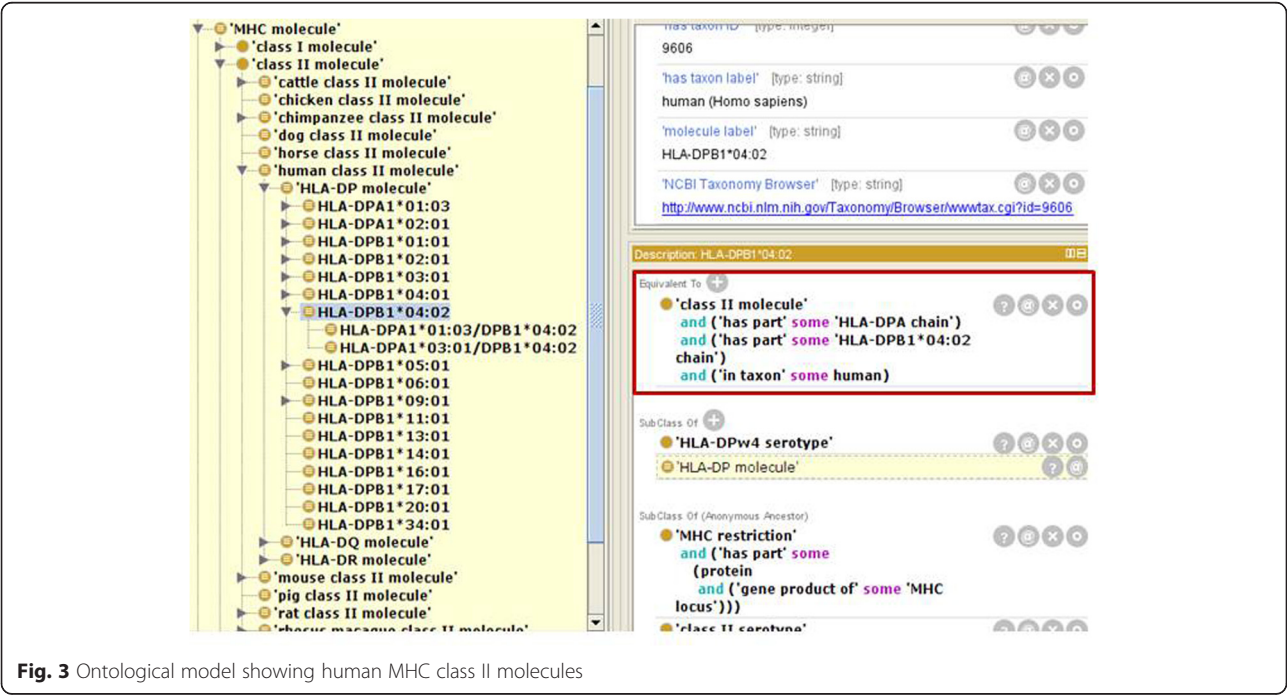
In MHC binding and elution assays, the exact MHC molecule studied is typically known; however this is often not the case for T cell assays. When a T cell responds to an epitope, the identity of the MHC molecule presenting the epitope may not be known at all, it may be narrowed down to a subset of all possible molecules or it may be exactly identified. In the context of T cell assays, the MHC restriction can be determined by the genetic background of the host, conditions of the experiment, or the biological process being measured; therefore we represent MHC molecules at a variety of levels and specify the rationale behind the determined restriction using evidence codes.

As shown in Fig. 4a, IEDB Evidence codes include “author statement” for cases where authors report previously defined restriction and “MHC ligand assay” used for MHC restriction established via an experiment that demonstrated the ability of the epitope to bind strongly to the MHC molecule or to have been eluted from that molecule. Figure 4b shows the metadata associated with this evidence code. “MHC binding prediction” is used when computer algorithms are used to predict the likelihood of an epitope to bind to a specific MHC molecule. In cases where authors analyze the MHC phenotype of a study population and conclude a

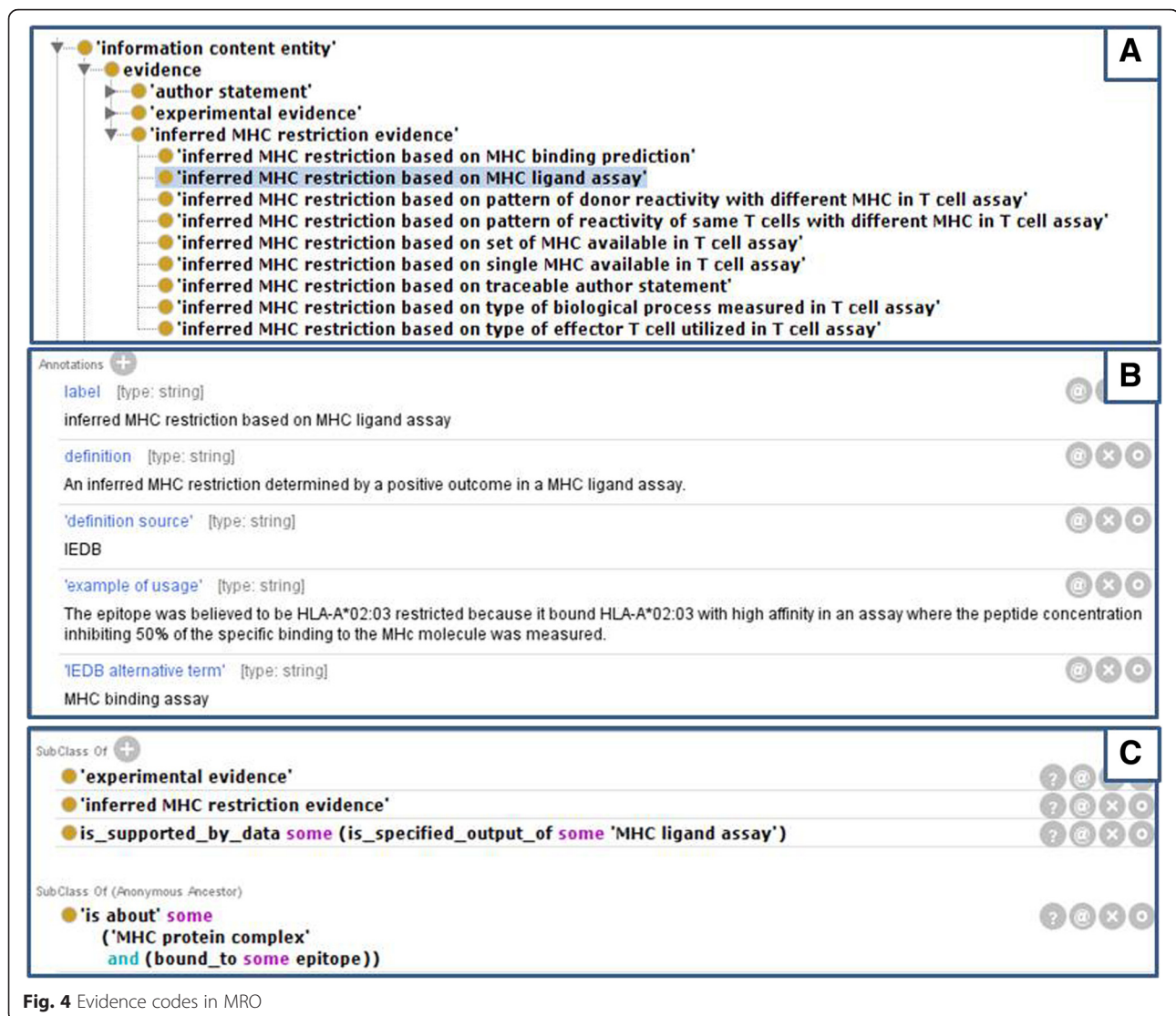


likely restriction based upon epitope recognition patterns among the subjects, “statistical association” is used as the evidence code. We use a set of evidence codes to communicate restriction shown by the response of T cells to the epitope: MHC complex. These include “Single MHC available” for cases where T cells respond to the epitope when only a single MHC molecule is available and “reactivity of same T cells with different MHC” is used when different APC expressing different MHC are used to narrow the potential restriction. The use of antibodies to block or purify subsets of MHC molecules typically determines

restriction to an imprecise level, such as HLA-DR and is conveyed by “set of MHC available.” When the T cells being studied are known to be CD8 or CD4 cells, the restriction can be deduced to be class I or class II, respectively, due to the known binding pattern of the molecules, as depicted in Fig. 1c. This case is communicated by the evidence code of “type of effector T cell.” Lastly, certain T cell responses can indicate the effector cell phenotype of CD8 or CD4, based upon known functions of the subsets and thus, class I or II restriction can be inferred and is noted by the evidence code of “biological process measured.” Figure 4c shows the







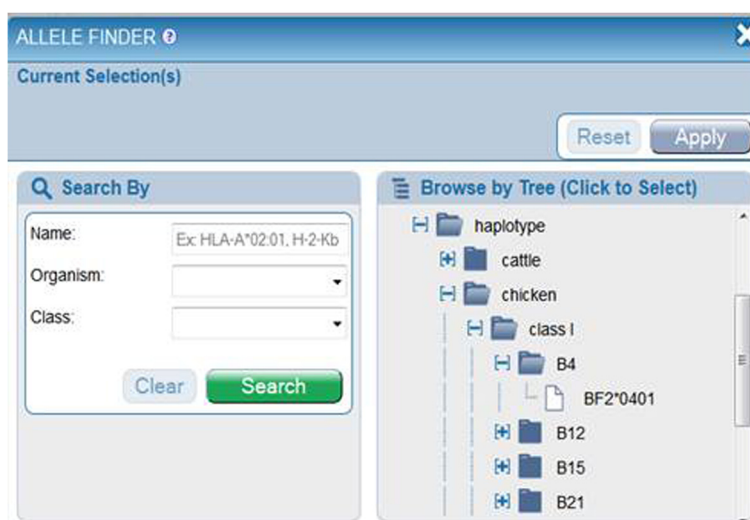
**Fig. 4** Evidence codes in MRO

modeling of these evidence codes in terms of the specific experiments, data transformations performed (using OBI terms), and the type of conclusion drawn. This work is being conducted in parallel with the general alignment of the Evidence Ontology (ECO) [14], which provides succinct codes for such types of evidence, with OBI, which can break down how such a code translates to specific experiments performed.

The IEDB MHC Allele Finder application, shown in Fig. 5, now allows users to browse data in different views. MHC molecules are first categorized into 'class I, class II or non-classical', and then further subdivided by species. Within each species, MHC molecules are organized by genetic locus. For select species, such as human, there are a large number of MHC molecules known and studied per genetic locus, thus sub-loci are

also used in order to present the data in a more user-friendly format. Each MHC molecule is presented under its locus, its haplotype, and/or its serotype, when available, all representing newly added functionalities. The haplotype the host species expresses is represented as immunologists often rely on the known haplotypes of research animals to narrow the potential MHC restriction. For example, when BALB/c (H2d) mice demonstrate a response to an epitope and the responding T cells are CD4+, the restricting MHC can be assumed to be one of the two MHC class II molecules of that haplotype, namely H2 IAd or IEEd.

The serotype of an MHC molecule, defined by antibody staining patterns, is relevant in immunology as this was the method of choice to identify MHC molecules until quite recently. In contrast to molecular definitions



**Fig. 5** IEDB's MHC Allele Finder, demonstrating chicken haplotypes

of MHC molecules based on their specific nucleotide or amino acid sequence, serotyping classifies MHC molecules based entirely on antibody binding patterns to the MHC molecule. These patterns are linked to the panel of antibodies used. Changing the antibody panel changes the serotype of a molecule. This can result in “serotype splits” where MHC molecules that were previously considered identical by one antibody panel, are later found to actually be two different molecules by a different antibody panel. To reflect this extrinsic nature of serotyping, we refer to serotypes as information entities rather than physical entities. Alternatively, the concept of serotype could also be modeled as collections of binding dispositions, but we chose what we thought was the simpler approach. MHC for all 16 species currently having MHC data in the IEDB are modeled to give users the ability to browse the tree in multiple ways and search IEDB data broadly, by entire MHC class, for example, or narrowly by a specific MHC protein chain. As new MHC molecules are encountered, they can be easily incorporated into this ontology.

## Conclusions

In conclusion, we formally represented MHC data building on established ontologies in order to represent MHC restrictions as required by immunologists. Accordingly, we modeled MHC molecules as a protein complex of two chains and established the relationships between the genes encoding these proteins, the haplotypes expressed by specific species, and the MHC classes. Traditional serotype information was also related to specific MHC molecules. Precise MHC restriction

was conveyed, as well as inferred MHC restriction and also the experimental evidence upon which the restriction was established. We will continue to formalize this work and will release a completed interoperable ontology later this year. Thus, MHC data in the IEDB is now presented to its users in a hierarchical format which simplifies searching the data and additionally instructs users on the inherent relationships between MHC genes and MHC restriction.

## Additional file

**Additional file 1: Goals and status of the MRO project.** (XLSX 16 kb)

## Abbreviations

MHC: Major histocompatibility complex; IEDB: The Immune Epitope Database; APC: Antigen presenting cell; HLA: Human leukocyte antigen; IMGT: ImMunoGeneTics; IPD: Immuno Polymorphism Database; MRO: MHC Restriction Ontology; BFO: Basic Formal Ontology; GO: Gene Ontology; PRO: Protein Ontology; OBI: Ontology for Biomedical Investigations; ECO: Evidence Ontology; OBO: The Open Biomedical Ontologies.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

RV, JAO and BP conceived of the ontology, and participated in its design and coordination and helped to draft the manuscript. ES prepared and analyzed MHC datasets. JS, JK, RLT, SE, JH, GWB, AS and BP provided expert guidance with relationship to specific MHC subsets to direct the development of the ontology. All authors read and approved the final manuscript.

## Acknowledgements

We wish to thank Kirsten Fischer Lindahl and Lutz Walter for their kind assistance with the mouse and rat MHC molecule nomenclatures, respectively. The Immune Epitope Database and Analysis Project is funded by the National Institutes of Health [HHSN272201200010C].

**Author details**

<sup>1</sup>La Jolla Institute for Allergy and Immunology, 9420 Athena Circle La Jolla, San Diego, California 92037, USA. <sup>2</sup>University of Cambridge, Trinity Ln, Cambridge CB2 1TN, UK. <sup>3</sup>Cornell University College of Veterinary Medicine, Ithaca, New York 14853-6401, USA. <sup>4</sup>The Pirbright Institute, Ash Rd, Woking GU24 0NF, UK. <sup>5</sup>The Babraham Institute, Cambridge CB22 3AT, UK.

Received: 29 September 2015 Accepted: 3 January 2016

Published online: 11 January 2016

**References**

1. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 2015; 43(Database issue):D405–12.
2. Sette A, Fleri W, Peters B, Sathiamurthy M, Bui HH, Wilson S. A roadmap for the immunomics of category A–C pathogens. *Immunity.* 2005;22(2):155–61.
3. DeLuca DS, Beisswanger E, Wermter J, Horn PA, Hahn U, Blasczyk R. MaHCO: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining. *Bioinformatics.* 2009;25(16):2064–70.
4. Overton JA, Dietze H, Essaid S, Osumi-Sutherland D, Mungall CJ. ROBOT: A command-line tool for ontology development. Lisbon, Portugal: 5th International Conference on Biomedical Ontology; 2015. July 29.
5. Simon J, Dos Santos M, Fielding J, Smith B. Formal ontology for natural language processing and the integration of biomedical databases. *Int J Med Inform.* 2006;75(3–4):224–31.
6. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Na Genet.* 2000;25(1):25–9.
7. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.* 2011;39:D539–45.
8. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. OBI consortium. Modeling biomedical experimental processes with OBI. *J Biomed Semantics.* 2010;1:S7.
9. Brush MH, Vasilevsky N, Torniai C, Johnson T, Shaffer C, Haendel MA. Developing a Reagent Application Ontology within the OBO Foundry Framework. Buffalo, NY: Proceedings of the International Conference on Biomedical Ontology; 2011.
10. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biol.* 2005;6(5):R46.
11. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2009;37:D5–15.
12. Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 2007;25:1251–5.
13. Glimm B, Horrocks I, Motik B, Stoilos G, Wang Z. HermiT: an OWL 2 reasoner. *J Automated Reasoning.* 2014;53(3):245–69.
14. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database.* 2014: 1–11.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

